The College Board

Research Notes

Office of Research and Development RN-09, April 2000

An Overview of Computer-Based Testing

The increased usage of computer technology in instruction has led educational institutions to look for ways to use this technology in testing. The age of the number-two pencil in standardized assessment is far from over, but computer-based testing (CBT) is becoming more popular. Because such a wide range of CBTs exists, educators who are trying to make decisions about utilizing CBTs are often unaware of the options available to them. To assist those educators, this paper provides a general overview of CBTs and distinguishes these systems on a single continuum.

Because different types of computerized tests exist and continue to emerge, the term of "computer-based testing" does not encompass all of the various models that may exist. As a result, test delivery model (TDM) is used to describe the variety of methods that exist in delivering tests to examinees. The criterion that is used to distinguish between the various TDMs is the extent to which the test is adaptive to an examinee's performance during the test session. As illustrated in Figure 1, the degree of adaptivity ranges from a linear to an adaptive test.

On one side of the continuum, there are linear tests that do not change in light of the examinee's performance. On the other side of this continuum, there are tests in which each item presented is dependent on the examinee's performance (i.e., adaptive). Therefore, not all tests that are computerized are adaptive to an individual's performance during the test. To distinguish this feature, adaptive tests are called computer-adaptive tests (CATs).

KEYWORDS:

Computer-Based Testing
Computer-Adaptive Testing
Test Delivery Models

The introduction of item response theory (IRT) has been responsible for the adaptive nature of tests administered by computer. IRT provides a method by which we can ascertain an examinee's proficiency from his or her performance on a set of items. Using IRT methodology, certain measurement properties are calculated beforehand for each item. Using these properties, the optimum item is selected and presented to the examinee after each response (see Hambleton & Swaminathan, 1985). IRT has contributed to the adaptive nature of TDMs through (1) the construction of item banks, (2) the use of a common scale for items and examinee characteristics, (3) item selection procedures, and (4) customizing the test to suit the specified purpose (Kingsbury & Houser, 1993). Even though these are areas that have been assisted by IRT, there are many issues that remain to be fully resolved.

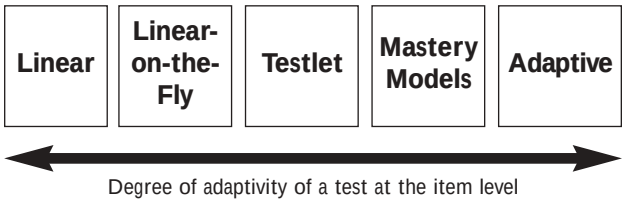

Figure 1. A continuum of test delivery models.

The distinguishing feature of TDMs is the extent to which the test is adaptive to an individual's performance. Potentially, there are numerous testing innovations that computer technology can provide that are slowly being realized (see Drasgow & Olson-Buchanan, 1999). Each of the general types of TDM will be discussed below. Please note that other test delivery models exist or could be developed but are not covered here.

TYPES OF TEST DELIVERY MODELS

Linear Tests

Linear tests, as defined earlier, are tests that are not adaptively administered. This type of test displays no adaptivity, and is thus on the far-left side of Figure 1. The term linear represents the sequential nature of the administration of the items on the tests. Specifically, the examinee is presented

with the first item, then the second, the third, etc. in a predetermined fashion. Due to the nonadaptive, predetermined nature of these tests, linear tests administered on the computer are also known as fixed-form tests.

Such tests were developed to reflect the same properties as paper-and-pencil tests but are administered on the computer. The test construction and psychometric methods employed in the development of paper-and-pencil tests are utilized in linear tests. Each item is presented to the examinee in a specified order, in the same manner as paper-and-pencil tests.

Such tests can be administered and scored on an as-needed basis. In linear test delivery models, the identical test forms are given to examinees, and examinees are allowed to review, revise, and omit items. Because these tests are developed in the same manner as paper-and-pencil tests, examinees are familiar with this type of testing. Thus, the linear test delivery models are easier to implement and explain. However, if the purpose of a test is compromised by the administration of identical forms, linear or fixed-form test delivery models may produce a security risk.

Like all other test delivery models, reporting can be automatically produced at the end of testing. This provides flexibility for testing directors at various institutions. Finally, data from testing in an electronic format may be imported into existing student database management systems.

Linear-on-the-Fly

In Linear-on-the-Fly Testing (LOFT), unique fixed-length tests are developed for each examinee. A unique form is assembled at the beginning of each test session to meet a target set of content and psychometric specifications. The items in this form are not dependent on the examinee's proficiency level; thus, these tests are not adaptive. Finally, a large number of items in the pool are needed in order to develop the unique forms.

LOFT is beneficial in testing programs that have concerns about item exposure and rigorous content ordering requirements. Therefore, the advantage of the LOFT delivery model over the linear model is improved security that comes from some randomization of items across forms. Finally, LOFT delivery models permit examinees to review, revise, and omit items.

Testlet

Testlets are a number of items that are considered a unit and administered together. Usually, testlets are constructed in advance, using prior knowledge of the difficulty of the items or knowledge of content experts. Testlets are assembled to allow the ordering of item difficulty or to meet content specifications, and are presented to examinees as units. Within testlets, examinees are permitted to review, revise, and omit items. Items within a testlet may be designed to deliver testlets based on equal difficulty, subject matter, or two- and multi-stage testing[1].

Mastery Models

These tests are developed to provide accurate information about mastery/non-mastery. These models have the goals of (1) covering the content domain and (2) making accurate mastery decisions. There are numerous models to implement mastery models, and their major advantage is efficiency. Efficiency is observed in that all examinees can be easily classified based on simple decision rules.

Adaptive (CAT)

Tests based on the CAT delivery model, utilizing IRT, present items depending on the performance of the examinee. The items that are presented have been pretested and item parameter estimates have been calculated. Using this information, examinees receive items that match the examinee's proficiency level at that time. Items are continually selected and presented until either the accuracy of the test score (i.e., standard error of measurement) reaches certain levels or, for some tests, the test ends after a specified number of items are administered (i.e., fixed-length CAT). The length for fixed-length tests is usually established after sufficient research has shown that the scores are reliable. The advantages of either type of CAT are (1) efficiency, (2) broad range of measurement, and (3) increased security (Ward, 1988). In terms of efficiency, 50 percent or greater reductions in

[1] The two- and multi-stage testing delivery system involves two or more stages that an examinee takes. Each stage involves the presentation of a testlet to the examinee. The first testlet or first stage is a testlet of average difficulty. Performance of the examinee on this stage determines the level of difficulty of the second testlet or stage. This type of delivery model may continue to present subsequent testlets based on the performance of the examinee on the previous testlet.

test length can be achieved while maintaining the accuracy of the measurements (Wainer, 1993).

Additionally, the adaptive nature of CAT allows measurement of a broad range of ability. Thus, examinees will receive test questions based on their ability levels. With conventional, non-adaptive tests, the items have been predetermined to work for a certain level of ability, usually at the average level.

Thirdly, with limited exposure to all items on the test, the CAT can increase the test security. However, there is some concern over exposure of items to the same examinees over time (Luecht, 1998; O'Neill, Lunz, & Thiede, 1998).

The disadvantages to CAT include the following: (1) the technical, both psychometric and computer, requirements, (2) the resources needed to develop the CAT (Ward, 1988), and (3) the user reaction to certain aspects of the CAT. In terms of the technical requirements, Ward (1988) indicated that the resources needed to build the item bank using a three-parameter model are quite costly. Ward suggested that at least 1,000 examinees should be used for preanalysis of items used for the item bank. The ability to review, revise, and omit items is usually not permissible in CAT. This may not be acceptable to certain examinees and certainly is contrary to past experiences of examinees who have taken paper-and-pencil tests.

Advantages and Disadvantages of CBTs

There are advantages and disadvantages in using any form of CBTs, as opposed to using paper-and-pencil assessments. The most salient advantage of a CBT is the immediacy in obtaining a score report. In addition, certain CBTs will provide a means of electronically transferring the results of testing to an existing database that is utilized by the institution for other functions.

The biggest disadvantage of CBTs is found in the costs relating to the start-up and maintenance of the computer environment (hardware, software, networking, and wiring) and the development and maintenance of sufficiently large item pools. The actual costs associated with a CBT will be dependent on each institution's existing technology resources. Institutions that are already utilizing computer technology will find the costs less than institutions that are not. These costs involve both financial and human resources.

Finally, as indicated above, the maintenance of the item pools will require a serious degree of commitment by the test publisher. The delivery model and the intended purpose of the test will determine the amount of resources that will be required. If security is an issue in a high-stakes environment, close monitoring with appropriate resources will be needed.

ARCHITECTURE OF A COMPUTER-BASED TESTING SYSTEM

Having mentioned the advantages and disadvantages of CBTs, this section of the paper will present an overview of the architecture of any CBT system. The organization of this discussion concerns the following areas: (1) item pools, (2) test algorithm, (3) delivery system, and (4) score reporting. While this is strictly an overview, references for detailed information are provided.

Item Pools

The core of any testing program is the items it comprises. The resources needed for quality items, especially for CAT, go beyond the initial development of items. There must be a system in place concerning the maintenance and renewal of the inventory of items. While most research focuses on exposure rates of items to maintain security (e.g., Stocking & Swanson, 1993), there is a growing interest to match the items to specific content (Kingsbury & Zara, 1989).

These types of constraints on the items have serious implications on the procedure used to maintain them. Current overviews of these approaches indicate that there must be a process of revising and recycling items over years of use (Way, Steffen, & Anderson, 1998) and a series of steps must exist to detect and correct evidence of item exposure (Davey & Nering, 1998).

Once the items have been developed, they must be pretested to estimate the item parameters using IRT methodology. The number of items should be large. The size will be based on such factors as the type of test delivery method, the measurement model used, the overall test length, the frequency of test administration dates, and security needs (or test stakes) for the test. However, the sheer number is not enough. The items must exist in sufficient number across the full range of per-

formance. This results in a great expense, since further item development may emerge after the initial pretesting. Parshall (1998) discusses the need for pretest items (and pretest examinees) across test delivery models, and some associated problems.

Compromises may be involved in developing items to cover the full range of performance. These compromises on the number of items include the following issues: (a) using a one- or two-parameter IRT model, (b) focusing the development of items near decision-points of performance (e.g., proficiency levels), and (c) selecting a test delivery model.

In building an item pool, the number of items in the pool must be larger than the number of items in the test. A 6:1 ratio of item pool size to test length is common in practice (Hambleton, Jones, & Rogers, 1993). In addition, the number of examinees should be large enough to estimate the item parameters during the pilot phase, and the number of items in the final pool should be relatively large in relation to the test length. If these issues are sacrificed, the stopping rule should be increased.

Test Algorithm

While the test items are the core of a CBT, the mechanism that presents these items to examinees is the test algorithm. The test delivery models discussed above lay the foundation for the development of a test algorithm. There are three components involved in any test algorithm: (1) where to start the test, (2) how to continue, and (3) when to end the test. The test algorithm based on a CAT utilizes item information to select the item for presentation. Depending on the performance of the examinee on the item presented, the test algorithm selects the next item for presentation. This continues until either a certain number of items have been answered or a certain level of confidence in the performance of the examinee is reached.

There is a growing number of item selection strategies (see Folk and Smith, 1998). Each of these strategies provides both advantages and disadvantages. Folk and Smith (1998) organize the issues surrounding the delivery of the test to examinees in the following areas: (1) item versus testlet delivery, (2) control of item exposure, (3) fixed-length versus variable-length testing and stopping rules, (4) item review and omits, and (5) multiple cut scores. The issues presented above have practical consequences to the examinee and testing organization and must be specified in the test algorithm. The decisions made in each area are compromises to the goals of the test with the available resources in administrating the test.

Delivery System

This has received some attention by the testing community, more by the information technology community, and the most by finance people. The capacity and speed of modern computers and networks will soon permit the execution of complicated algorithms and use of complex items. However, more research is needed to develop the appropriate psychometric models.

Additional research examining the interaction between the computer and the examinees is needed. For example, in presenting a long reading passage as part of the test, what is the consequence of having the examinee scroll through the text versus providing the reading passage in a split-screen format? In addition, when mathematics questions are given, should additional space be provided with paper and pencils for the examinees to work their answers? The design of the examinee's space during testing has also not received sufficient attention.

Score Reporting

This aspect of CBTs involves both the scoring mechanism and the reporting of the results to the examinees. As mentioned earlier, one of the advantages of using a CBT is that scores may be provided immediately after testing.

In CATs, scoring is intertwined with IRT. However, there are numerous efforts to make score reporting meaningful and easier to interpret by the examinees (see Dodd & Fitzpatrick, 1998). These efforts include translating number correct scoring into IRT ability (e.g., Yen, 1984). Other efforts have examined the use of testlets in similar scoring approaches (Schnipke & Rees, 1997; Thissen, 1998). The efforts of the research have been focused on addressing the comparability of number correct scoring with item patterns. However, the results of this work should include the test consumers' perceptions and understanding.

While each area above has research to support it, there is not a prescription of what approach will work in any given situation.

FUTURE DIRECTIONS

Future applications taking advantage of modern computer technology (e.g., the Internet, sound recording, etc.) will only increase the use of computer-based tests. Test developers can utilize technology to permit more authentic and efficiently delivered stimuli. For example, speech and non-speech sound can make tests more appealing to the examinees (Parshall, 1999).

Utilizing today's technology, there are limitless possibilities for the types of tests that can be developed. (For additional examples of innovative computerized assessment see Drasgow and Olson-Buchanan, 1999.) As Leucht and Clauser (1998) have suggested, we can create a virtual world and submerge the examinee in it; however, we do not know how such conditions translate into our knowledge about how to score and what the score means. Thus, in moving to more CBTs, caution must be exercised.

The author is Thanos Patelis, assistant research scientist at the College Board.

REFERENCES

Davey, T. & Nering, M. (September 1998). *Controlling item exposure & maintaining item security.* Invited paper presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

Dodd, B. G. & Fitzpatrick, S. J. (September 1998). *Alternatives for scoring computer-based tests.* Invited paper, presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

Drasgow, F. & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Folk, V. G. & Smith, R. L. (September 1998). *Models for delivery of computer-based tests.* Invited paper presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143-155.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory.* Boston: Kluwer-Nijhoff Publishing.

Kingsbury, G. G. & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice, 12*(1), 21-27, 39.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-386.

Luecht, R. M. (April 1998). *A framework for exploring and controlling risks associated with test item exposure over time.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Luecht, R. M. & Clauser, B. E. (1998). *Test models for complex computer-based testing.* Invited paper presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

O'Neill, T., Lunz, M. A., & Thiede, K. (April 1998). *The impact of item exposure on repeat candidate performance.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Parshall, C. G. (1998). *Items development and pretesting in a computer-based testing environment.* Invited paper presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

Parshall, C. G. (April 1999). *Measuring more through the use of speech and non-speech sound.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Schnipke, D. L. & Reese, L. M. (March 1997). *A comparison of testlet-based test designs for computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Stocking, M. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Thissen, D. (April 1998). *Scaled scores for CATs based on linear combinations of testlet scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*(1), 15-20.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning,* 2, 271-282.

Way, W. D., Steffen, M., & Anderson, G. S. (September 1998). *Developing, maintaining, and renewing the item inventory to support computer-based testing.* Invited paper presented at the Educational Testing Service's Computer Based Testing Colloquium: Building the Foundation for Future Assessments Colloquium, Philadelphia.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93-111.

For more information or additional copies of this report, please write to: Office of Research, The College Board, 45 Columbus Avenue, New York, NY 10023-6992, or contact us by e-mail at: research@collegeboard.org, or visit our Web site at: www.collegeboard.com.

4/00
988458